Resumen

Aunque la industria musical siempre ha buscado formas de interpretar datos para desarrollar estrategias que impulsen a artistas y lanzamientos, el surgimiento del streaming ha transformado la música en un elemento cotidiano y esencial en la vida de millones de personas. Esto ha generado una cantidad exponencial de datos disponibles para análisis. Sin embargo, la abundancia de variables hace que comprender qué características musicales contribuyen al éxito de una canción siga siendo un desafío tanto para los artistas como para las plataformas de streaming. Este proyecto aborda ese reto mediante el análisis de una base de datos de canciones disponibles en Spotify, que abarca el período de 1998 a 2020. El objetivo principal es identificar patrones y agrupar canciones que compartan características como género, bailabilidad, energía y popularidad, para entender cómo estos elementos están relacionados con el éxito.

La importancia de este trabajo radica en la aplicación de técnicas de aprendizaje no supervisado, como el clustering y la reducción de dimensionalidad, para descubrir patrones en los datos musicales. Estas técnicas permiten segmentar canciones en grupos basados en características comunes, lo que puede proporcionar información valiosa para los artistas al momento de crear nuevas canciones. Con esta información, los artistas podrían enfocar su producción musical en aquellas características que incrementan las posibilidades de éxito, algo especialmente relevante en una industria tan competitiva y cambiante.

Para las plataformas de streaming, como Spotify, el análisis de estos patrones y grupos no solo puede mejorar las recomendaciones personalizadas, optimizando la experiencia del usuario, sino también facilitar la creación de listas de reproducción más atractivas y eficaces. Además, los insights obtenidos podrían servir para diseñar estrategias de marketing más precisas, dirigidas a audiencias específicas basadas en sus preferencias musicales.

Para abordar este reto, se revisaron antecedentes académicos sobre el análisis de datos musicales. Se encontraron proyectos que han implementado técnicas como el PCA y algoritmos de clustering, como K-means y DBSCAN, para agrupar canciones según sus características, sin necesidad de etiquetas de género. También se identificaron estudios que usaron la información obtenida de los clusters para clasificar canciones en géneros a partir de las letras, sin etiquetas previas.

Finalmente, el equipo revisó la base de datos a trabajar, comprendió el significado de las variables disponibles, y analizó la estadística descriptiva de los datos para desarrollar la propuesta metodológica. Esta incluye la extracción y limpieza de datos, seguida por la reducción de dimensionalidad, la elección e implementación de métodos de clustering, y un análisis final de los resultados obtenidos.

Introducción

En la era digital, la música se ha convertido en una parte integral de la vida cotidiana, y plataformas como Spotify han transformado la forma en que consumimos y descubrimos música. Se obtiene una vasta biblioteca de canciones que abarca desde 1998 hasta 2020, para explorar patrones y características de la música disponible en Spotify. La plataforma recopila una variedad de información sobre las canciones, incluyendo atributos como el género, la bailabilidad, la energía y la popularidad, así como las emociones que transmiten. Este proyecto se centra en el análisis de esta base de datos con el objetivo de descubrir patrones y grupos dentro de la música, lo que puede proporcionar valiosas percepciones sobre cómo ciertas características musicales se relacionan con el éxito de una canción.

El problema central que se aborda en este proyecto es la identificación de grupos de canciones que comparten características similares, tales como género, bailabilidad, energía y popularidad. La pregunta específica que se busca responder es ¿Existen grupos de canciones que comparten características similares, como género, bailabilidad, energía, o popularidad, que podrían ayudar a predecir su éxito?

La motivación para resolver esta pregunta radica en la posibilidad de aplicar técnicas de aprendizaje no supervisado, específicamente el clustering, para revelar estructuras ocultas en los datos musicales. El clustering es un área del aprendizaje no supervisado que se enfoca en agrupar datos en subconjuntos, de manera que los elementos dentro de un mismo grupo sean más similares entre sí en comparación con otros grupos. Identificar patrones y grupos de canciones con características

comunes puede tener implicaciones significativas tanto para los artistas como para las plataformas de streaming(En este caso Spotify). Para los artistas, comprender qué características musicales tienden a ser más exitosas puede informar la creación de nuevas canciones y álbumes. Para las plataformas de streaming, como Spotify, el análisis de estos patrones puede mejorar las recomendaciones personalizadas, optimizar las listas de reproducción y diseñar estrategias de marketing más efectivas. Además, al descubrir grupos de canciones con características similares, los datos pueden ser utilizados para predecir el éxito de nuevas canciones, proporcionando así una ventaja competitiva en un mercado musical altamente dinámico. Este análisis no solo contribuirá a una comprensión más profunda del impacto de las características musicales en el éxito, sino que también permitirá una mejor adaptación a las tendencias emergentes en la industria de la música.

Revisión preliminar de antecedentes

El análisis de grandes volúmenes de datos musicales, como los proporcionados por plataformas de streaming, ha impulsado el desarrollo de técnicas avanzadas de aprendizaje no supervisado. Estas técnicas han demostrado ser fundamentales para identificar patrones ocultos y agrupar canciones basadas en características como el género, la energía y la popularidad. Este enfoque se alinea con las investigaciones recientes en análisis musical, que buscan aprovechar estas metodologías para mejorar la predicción del éxito de canciones y optimizar las recomendaciones en plataformas como Spotify.

Un aspecto clave del uso de clustering en este contexto es la capacidad de identificar las características que diferencian las canciones populares de las impopulares. Estudios recientes han demostrado que el análisis de atributos como la bailabilidad, la energía y la popularidad permite a los algoritmos de clustering crear grupos de canciones que comparten estos atributos, revelando así las características comunes entre las canciones exitosas. Por ejemplo, el uso de **K-means** y **DBSCAN** ha sido ampliamente explorado en la literatura, destacando su capacidad para identificar subgéneros musicales y otras estructuras latentes dentro de conjuntos de datos musicales. En particular, el artículo "*Supervised Learning and Unsupervised Learning on Music Data*" explora cómo estas técnicas permiten agrupar canciones sin la necesidad de etiquetas predefinidas, lo que es fundamental en escenarios donde los datos etiquetados son limitados.

Además, el **PCA (Análisis de Componentes Principales)** se utiliza como una herramienta para simplificar los datos musicales antes de aplicar técnicas de clustering, lo que facilita la identificación de patrones significativos. Este enfoque ha sido discutido en el artículo "*A Study on Music Genre Classification using Machine Learning* ", donde se destaca la importancia de la reducción de dimensionalidad para mejorar la eficiencia de los algoritmos no supervisados.

El artículo "*Clustering Music by Genres Using Supervised and Unsupervised Algorithm*" también enfatiza el uso de K-means y PCA en la clasificación de géneros musicales. Los autores encontraron que, al aplicar PCA antes de realizar el clustering, se mejora la precisión en la identificación de géneros, especialmente cuando se manejan múltiples géneros musicales. Esto subraya la relevancia de la reducción de dimensionalidad como paso preliminar para mejorar el rendimiento de los algoritmos no supervisados en la clasificación de música.

Por otro lado, el artículo "*Unsupervised Learning for Music Genre Classification of Song Lyrics*" investiga la clasificación de géneros musicales basándose en las letras de las canciones utilizando métodos como **K-means** y **Gaussian Mixture Models (GMM)**. Este estudio revela que el uso de datos textuales, como las letras de las canciones, combinado con técnicas de clustering, puede proporcionar una visión complementaria a la clasificación basada en características acústicas. Además, los autores exploraron la eficacia de técnicas avanzadas de procesamiento de lenguaje natural (NLP), como **Doc2Vec**, para mejorar la agrupación de canciones por género.

En comparación con los métodos supervisados, como las redes neuronales convolucionales (CNNs), que también se han utilizado para la clasificación musical, las técnicas no supervisadas ofrecen una flexibilidad crucial. El artículo "*Music Genre Classification using Machine Learning: A Comparative Study*" subraya que, aunque las CNNs son eficaces, las técnicas de clustering no supervisado siguen siendo esenciales para explorar grandes volúmenes de datos sin depender de etiquetas explícitas.

A pesar de los desafíos, el aprendizaje no supervisado abre nuevas oportunidades para la exploración de la música digital. El enfoque de este proyecto se basa en estas técnicas avanzadas para descubrir patrones y agrupar canciones en función de

características como el género, la energía y la popularidad, lo que permitirá predecir mejor el éxito de nuevas canciones y comprender las características que las diferencian de las canciones menos populares.

Este enfoque no solo se apoya en la literatura revisada, sino que también busca extender el conocimiento en el campo, al aplicar estas técnicas a un conjunto de datos amplio y diverso, como el de Spotify, proporcionando insights valiosos tanto para artistas como para plataformas de streaming.

Descripción detallada de los datos.

Para el trabajo se cuenta con una base de datos de la plataforma Spotify, la cual fue procesada para aplicar los métodos de reducción de dimensionalidad y agrupamiento por clustering. Fuente de datos: https://www.kaggle.com/code/varunsaikanuri/spotify-data-visualization/comments

La base de datos se pasó a un Excel que contiene información asociada a la música escuchada en la plataforma Spotify en un rango de tiempo que va desde 1998 hasta 2020. En el pretratamiento los nombres de las variables fueron traducidos al español y se identificaron 12 géneros musicales, a continuación, en la tabla 1, se encuentra una breve descripción de las variables presentes en la base de datos, la cual está compuesta actualmente por 14 variables numéricas y 5 categóricas. En la descripción se ve como cada variable puede contribuir a la caracterización.

Tabla 1 Variables

Variable	Tipo de variable	Descripción
Artista	categórica	nombre del artista
Canción	categórica	Nombre del track o nombre de la canción
Duracion_ms	numérica	Duración en milisegundos de la pista
Explicito	categórica	La letra o el contenido de una canción o un vídeo musical contienen criterios que podrían considerarse ofensivos o inadecuados para los niños.
Anio	numérica	Año de lanzamiento de la pista.
Popularidad	numérica	cuanto mayor sea el valor, más popular será la canción.
Bailable	numérica	la bailabilidad describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general. Un valor de 0,0 es el menos bailable y 1,0 es el más bailable.
Energía	numérica	La energía es una medida de 0,0 a 1,0 y representa una medida perceptual de intensidad y actividad.
Clave (variable de codificación de las notas musicales)	numérica	la clave en la que se encuentra la pista. Los números enteros se asignan a los tonos utilizando la notación estándar de clase de tono. P.ej. 0 = C, 1 = C♯/D♭, 2 = D, y así sucesivamente. Si no se detectó ninguna clave, el valor es -1. Corresponden a las notas musicales.
Volumen	numérica	el volumen general de una pista en decibeles (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar el volumen relativo de las pistas. El volumen es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db.
Modo	numérica	Modo indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. El mayor está representado por 1 y el menor es 0.
locuacidad	numérica	el habla detecta la presencia de palabras habladas en una pista. Los valores superiores a 0,66 = pistas que probablemente estén compuestas exclusivamente de palabras habladas. Los valores entre 0,33 y 0,66 = pistas que pueden contener música y voz, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores inferiores a 0,33 probablemente representen música y otras pistas que no sean de voz.

Acusticidad	numérica	una medida de confianza de 0,0 a 1,0 sobre si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.
Instrumentalidad	numérica	predice si una pista no contiene voces. Cuanto más cerca esté el valor de instrumentalidad de 1,0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0,5 pretenden representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0.
Vivisidad	numérica	Detecta la presencia de una audiencia en la grabación. Los valores de vivisidad más altos representan una mayor probabilidad de que la pista se haya interpretado en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté activa.
Valencia emocional	numérica	Una medida de 0,0 a 1,0 que describe la positividad musical transmitida por una pista. Las pistas con valencia alta suenan más positivas (por ejemplo, felices, alegres, eufóricas), mientras que las pistas con valencia baja suenan más negativas (por ejemplo, tristes, deprimidas, enojadas).
Tempo	numérica	El tempo general estimado de una pista en pulsaciones por minuto (BPM).
Género	categórica	Género musical de la pista.
Categoría_instrumentalidad	numérica	Es un campo calculado que implementa el campo instrumentalidad, si el valor es mayor o igual a 0.3 se cataloga como canción instrumental, en caso contrario es canción vocal

Para las variables numéricas se ha analizado la estadística descriptiva, encontrándose variables con rangos de valores diversos, se puede resaltar la variable popularidad que es clave para nuestro proyecto, presenta valores entre 0 y 89, con un promedio de 59.6 y el 50% de sus datos por debajo de 56, lo cual indica que entre las canciones se encuentra una gran variabilidad para este criterio, las canciones tienen una popularidad baja y media, mientras que popularidades superiores al 73% sólo se ven en un bajo porcentaje de canciones aproximadamente el 25%.Como se mencionó anteriormente, se confirma que la base de datos contiene canciones de entre 1998 y 2020. Luego del verificar y eliminar valores nulos y duplicados, la base de datos tiene un total de 1925 registros de canciones. Los resultados estadística descriptiva se encuentran a continuación:

1 Estadística descriptiva para las variables numéricas

	duracion_ms	anio	popularidad	bailable	energia
count	1925.000000	1925.000000	1925.000000	1925.000000	1925.000000
mean	228668.634286	2009.487792	59.657143	0.607280	0.656745
std	39293.681728	5.874863	21.441402	0.226832	0.245218
min	113000.000000	1998.000000	0.000000	0.004000	0.004000
25%	203520.000000	2004.000000	56.000000	0.535000	0.572000
50%	223266.000000	2010.000000	65.000000	0.661000	0.716000
75%	248066.000000	2015.000000	73.000000	0.754000	0.833000
max	484146.000000	2020.000000	89.000000	0.975000	0.999000

	clave	volumen	modo	locuacidad	acusticidad
count	1925.000000	1925.000000	1925.000000	1925.000000	1925.000000
mean	5.364156	-5.031660	0.553766	0.376673	0.356154
std	3.616471	2.344846	0.497230	0.233340	0.256533
min	0.000000	-20.514000	0.000000	0.001000	0.000002
25%	2.000000	-6.313000	0.000000	0.202000	0.154000
50%	6.000000	-5.074000	1.000000	0.346000	0.285000
75%	8.000000	-3.832000	1.000000	0.504000	0.523000
max	11.000000	-0.036000	1.000000	0.998000	0.996000

	instrumentalidad	vivisidad	valencia emocional	tempo
count	1925.000000	1925.000000	1925.000000	1925.000000
mean	0.015522	0.353336	0.505283	108.280502
std	0.088943	0.283536	0.258660	42.825802
min	0.000000	0.001000	0.001000	0.760000
25%	0.000000	0.123000	0.321000	93.013000
50%	0.000000	0.247000	0.527000	116.043000
75%	0.000071	0.577000	0.715000	131.012000
max	0.985000	0.998000	0.973000	210.851000

Para el caso de las variables categóricas de menor cardinalidad (Explícito, nota clave y genero) se realizaron histogramas para evaluar las categorías predominantes, en el caso de la variable explícito la mayoría de las canciones no tienen contenido que pueda ser sensible para los niños, sin embargo, la cantidad de canciones que sí lo tienen no es despreciable. Para la variable nota_clave, predomina la nota “DO” en las canciones y la menos usual es la nota “RE”, las demás notas musicales tienen una frecuencia moderada en la base de datos de las canciones evaluadas. Finalmente, en la variable género, se encontraron 12, predominando el pop, hip-hop, rock y dance, mientras que los demás géneros no se encuentran muy representados en la base de datos, sus frecuencias son muy bajas respecto a los 4 predominantes. En los diagramas de torta podemos observar también el marcado predominio de algunas de las clases de los datos en las variables explícito donde falso representa el 72.4% y también para el género musical donde pop representa el 46.8% ver anexos y notebook.

Para el caso de variables como canción y artista, se encontró una alta cardinalidad. Las canciones tienen 1879 clases, donde la canción "Sorry" es la que más se repite en la base de datos con 4 entradas, seguida de la canción "Higher" con 3 entradas, la mayoría de las canciones (1838) tienen solo 1 entrada en la base de datos, debido a estas características para los siguientes pasos del proyecto se evaluará la pertinencia de incluir esta variable para el propósito de nuestro análisis. En el caso de los artistas, se encuentran 835 artistas diferentes en la base de datos con "Rihana" liderando en frecuencia con 23 entradas que representan aproximadamente el 1.19% de los datos de la base, seguida de "Drake" (21), los artistas con menor presencia en la base de datos representan cada uno aproximadamente 0.05% de la base de datos y tienen 1 canción presente. Para esta variable predominan artistas que aparecen en la base de datos con un máximo de 5 veces, entre ellos los que aparecen máximo 1 vez son 506 artistas de los 835. Para observar mejore esto se evaluó el número de artistas en 4 rangos de frecuencia y se encontraron 746 (0-5), 62 (5-10), 23 (10-20) y 4(20-24).

Propuesta metodológica

Con el objetivo planteado en la pregunta problema, buscamos usar las herramientas que nos permitan identificar las variables de nuestra base de datos para diferenciación de canciones populares de las que no lo son, esto mediante las técnicas de aprendizaje no supervisado. Se busca segmentar las canciones en clusters que se basen en las características de las canciones, dadas por las variables seleccionadas y que a través de esto podamos ver como estos conjuntos de variables numéricas y categóricas se combinan para dar ideas acerca de la popularidad de las canciones.

La metodología para abordar el proyecto consiste en varios pasos:

1. Exploración y limpieza de datos: en este punto ya se ha adelantado eliminación de los duplicados y de los datos faltantes, renombrar las variables al español, posteriormente se planea seleccionar las variables de acuerdo con lo encontrado en la primera fase de exploración (sección "Descripción detallada de datos") y estandarizar estas variables para poder proceder con el análisis. Para las variables categóricas se evaluarán técnicas más pertinentes de codificación como One-Hot Encoding ó frecuency encoding, ya que algunas variables como artista o nota clave tienen una cardinalidad mayor que en el caso de la variable explícito, por ejemplo.
2. Reducir la dimensionalidad: esto se buscará mediante la aplicación de PCA, que permite esto conservando la informacion relevante presente en nuestra base de datos. Se probará si es adecuado un valor entre 80-95% de explicación de la varianza para escoger el número de componentes principales adecuados.
3. Clustering: Esto permitirá cumplir con el objetivo de agrupar y segmentar las canciones de acuerdo con sus características comunes. Se probará el algoritmo K-Means para la obtención de los clusters, mediante el método de codo y el coeficiente de Silhouette se analizará el número de clusters pertinente. Se pretende incorporar en el proyecto el uso de DBSCAN para el análisis ya que puede darnos un valioso complemento para manejar el "ruido" ó los posibles valores atípicos.
4. Análisis de la información: con los clusterings obtenidos, se tendrá la información de interés para poder evaluar cuales son los grupos relacionados con una mayor o menor popularidad y tener una idea de cómo las variables evaluadas se comportan en estos clusters y dan fuerza a la popularidad de la canción. Analizaremos si los resultados obtenidos tienen validez o coherencia con lo esperado. Al final los resultados obtenidos pueden ser además de lo mencionado en los otros apartados, usado en Spotify para crear listas de reproducción de éxitos de un periodo de tiempo determinado y recomendar las canciones que se esperan de mayor popularidad, además dar informacion a otros sectores de la industria sobre las características que hacen exitosa una canción que se conservan a través de las épocas. El proyecto al incluir información de muchas variables puede ser de gran utilidad.
5. Paralelamente a nuestro planteamiento metodológico, se explorarán otros algoritmos que por su robustez también a los outliers y por la complejidad que pueden tener la diversidad de variables pueden ser utilidad: k-medioides y para la reducción de dimensionalidad se probará SVD.

Para el fin del proyecto se puede plantear que las canciones más populares sobrepasen 60% en este parámetro.

Bibliografía

Liu, C., & Chao, Z. Supervised learning and unsupervised learning on music data with different genres. In Proceedings of the 2021 IEEE 7th International Conference on Big Data Intelligence and Computing (DataCom), 2021, pp. 7-12. DOI: 10.1109/DataCom53700.2021.00008.

Available at: https://www.researchgate.net/publication/359563242_Supervised_learning_and_unsupervised_learning_on_music_data_with_different_genres.

Kim, K., Yun, W., & Kim, R. Clustering Music by Genres Using Supervised and Unsupervised Algorithms. 2015. Available at: https://www.semanticscholar.org/paper/Clustering-Music-by-Genres-Using-Supervised-and-Kim-Yun/0d93faaeef4055a71137eaa1ca7261cbd4dfe172.

Poonia, S., Verma, C., & Malik, N. Music Genre Classification using Machine Learning: A Comparative Study. Journal Name, 2022, 13, 15-21. Available at: https://www.researchgate.net/publication/362619781_Music_Genre_Classification_using_Machine_Learning_A_Comparative_Study.

Aggarwal, M., Nair, V., & Sun, T. Unsupervised Learning for Music Genre Classification of Song Lyrics. CS 221, Fall 2018. Available at: http://web.stanford.edu/~manav/CS%20221%20Final%20Paper.

Ghosh, P., Mahapatra, S., Jana, S., & Jha, R. A Study on Music Genre Classification using Machine Learning. International Journal of Engineering Business and Social Science, 2023, 1(04), 308-320. DOI: 10.58451/ijebss.v1i04.55. Available at: https://www.researchgate.net/publication/370546962_A_Study_on_Music_Genre_Classification_using_Machine_Learning.

Anexos

1.Figuras referenciadas